Over-representation analysis (ORA) practical lab :g:Profiler

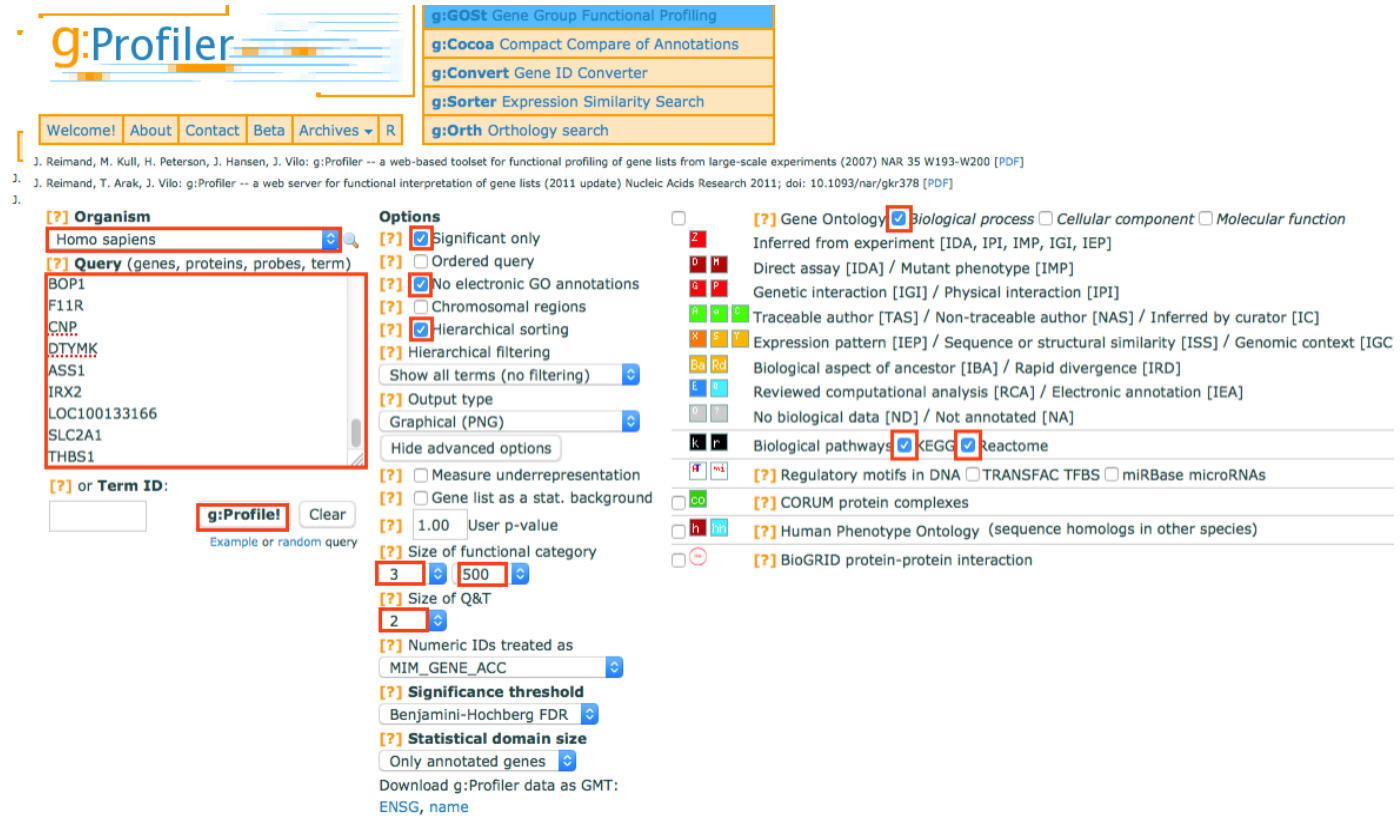
The data set used for this practical lab contains transcriptomics data obtained from MCF7 cells, a human breast cancer line, treated or non treated with estradiol. The cells were treated with estradiol for 12, 24 or 48 hours. Total RNA extracted from the cells was amplified, labeled and hybridized to Affymetrix GeneChip U133 Plus 2.0 microarrays. The data are available in the Gene Expression Omnibus (GEO) repository under the accession number GSE11352 (PMID: 17542648).

For this exercise, we are going to use a list of 428 genes that are differentially expressed in the MCF7 cells treated with estradiol for 24hr compared to the control samples. Our goal is to perform gene-set enrichment on this list using the g:Profiler tool and to explore the results. The Gene Ontology Biological Process, the KEGG and Reactome are going to be used as the pathway databases. g:Profiler uses a Fisher's exact test to calculate the significance of the gene-set enrichment.
Before starting this exercise, download the required file:

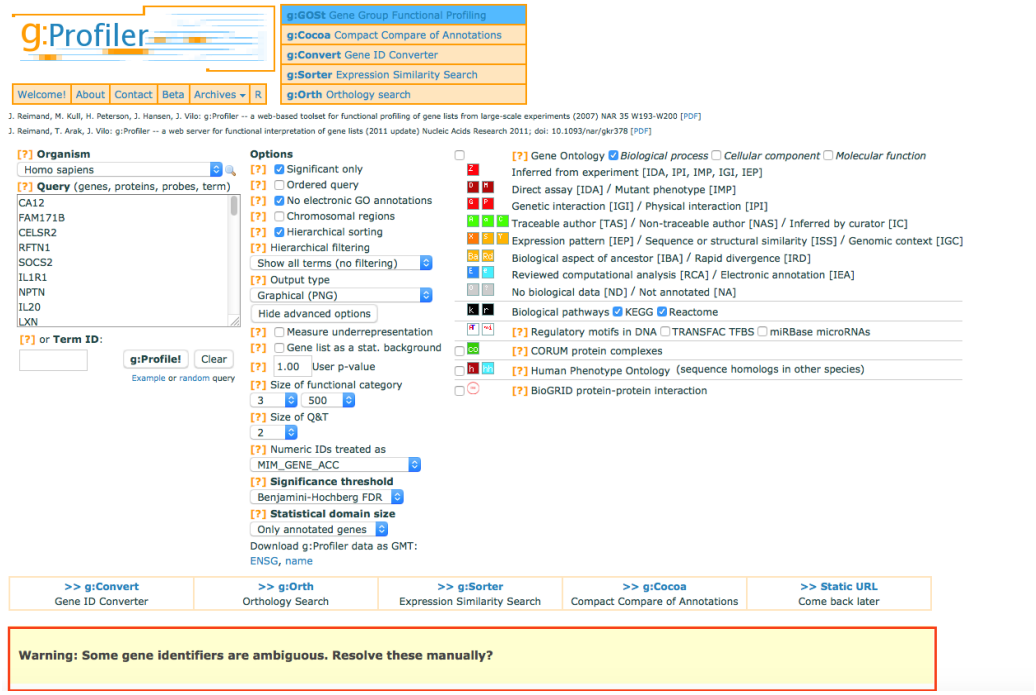
- 24hr_topgenes.txt .

Step	Action	Check
1	Go to g:Profiler 's homepage **at http://biit.cs.ut.ee/gprofiler/**	
2	Ensure that **'Organism'** is set to **'Homo sapiens'**	
3	In the '**Query'** box, **copy and paste** the 428 genes listed in the file 24hr_topgenes.txt	
4	Select the following 'Options' by checking the corresponding boxes: ● '**Significant only**' ● '**No electronic GO annotations**' ● '**Hierarchical sorting**'	
5	Select the following gene-set databases by checking the corresponding boxes: ● '**Gene Ontology**' '**Biological Process**' ● '**Biological pathways**' '**KEGG**' and '**Reactome**'	
6	Click on **'Show advanced options'**.	
7	Select the following 'advanced options': ● '**Size of functional category**' : 3 (min) and 500 (max) ● '**Size of Q&T'** : min of 2	
8	Click on the '**g:Profile!'** button.	
9	Click on the warning message **'Some gene identifiers are ambiguous. resolve these manually?'** Select the first *ENSEMBL ID* for each gene and click on **'Resubmit query'.**	
10	**Explore the results.** Which term has the best corrected p-value? Which genes in our list are included in this term?	
11	If time permits, play with input parameters, e.g. add **'TRANSFAC TFBS'** and **'miRBase microRNAs'** databases, rerun the query by clicking on the **'g:Profile!'** button and explore the new results.	

EXERCISE 2: Steps 1- 8



EXERCISE 2: Step 9



EXERCISE 2: Step 9 (continued)

Warning: Some gene identifiers are ambiguous. Resolve these manually?

Attempt to automatically resolve symbols using a namespace (percentage of ambiguous symbols resolved in brackets):

- VEGA_GENE (12%)
- HGNC (12%)

ARHGAP8
- *ENSG00000241484* (ARHGAP8, 9 GO annot.) - Rho GTPase activating protein 8 [Source:HGNC Symbol;Acc:HGNC:677]
- *ENSG00000248405* (PRR5-ARHGAP8, 9 GO annot.) - PRR5-ARHGAP8 readthrough [Source:HGNC Symbol;Acc:HGNC:34512]
- Ignore this gene

BOLA2
- *ENSG00000169627* (BOLA2B, 1 GO annot.) - bolA family member 2B [Source:HGNC Symbol;Acc:HGNC:32479]
- *ENSG00000183336* (BOLA2, 1 GO annot.) - bolA family member 2 [Source:HGNC Symbol;Acc:HGNC:29488]
- Ignore this gene

GPR89B
- *ENSG00000117262* (GPR89A, 14 GO annot.) - G protein-coupled receptor 89A [Source:HGNC Symbol;Acc:HGNC:31984]
- *ENSG00000188092* (GPR89B, 14 GO annot.) - G protein-coupled receptor 89B [Source:HGNC Symbol;Acc:HGNC:13840]
- Ignore this gene

MRPS17
- *ENSG00000239789* (MRPS17, 12 GO annot.) - mitochondrial ribosomal protein S17 [Source:HGNC Symbol;Acc:HGNC:14047]
- *ENSG00000249773* (MRPS17, 6 GO annot.) - 28S ribosomal protein S17, mitochondrial {ECO:0000313|Ensembl:ENSP00000390331}; HCG1984214, isoform CRA_a {ECO:0000313|E...
- Ignore this gene

PRICKLE4
- *ENSG00000124593* (PRICKLE4, 4 GO annot.) - prickle homolog 4 (Drosophila) [Source:HGNC Symbol;Acc:HGNC:16805]
- *ENSG00000278224* (PRICKLE4, 4 GO annot.) - prickle homolog 4 (Drosophila) [Source:HGNC Symbol;Acc:HGNC:16805]
- Ignore this gene

SERPINA3
- *ENSG00000273259* (SERPINA3, 14 GO annot.) - serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3 [Source:HGNC Symbol;Acc:HGNC:16]
- *ENSG00000196136* (SERPINA3, 14 GO annot.) - serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3 [Source:HGNC Symbol;Acc:HGNC:16]
- Ignore this gene

SGK3
- *ENSG00000104205* (SGK3, 23 GO annot.) - serum/glucocorticoid regulated kinase family, member 3 [Source:HGNC Symbol;Acc:HGNC:10812]
- *ENSG00000270024* (C8ORF44-SGK3, 23 GO annot.) - C8orf44-SGK3 readthrough [Source:HGNC Symbol;Acc:HGNC:48354]
- Ignore this gene

TXNDC5
- *ENSG00000259040* (BLOC1S5-TXNDC5, 5 GO annot.) - BLOC1S5-TXNDC5 readthrough (NMD candidate) [Source:HGNC Symbol;Acc:HGNC:42001]
- *ENSG00000239264* (TXNDC5, 14 GO annot.) - thioredoxin domain containing 5 (endoplasmic reticulum) [Source:HGNC Symbol;Acc:HGNC:21073]
- Ignore this gene

Resubmit query

EXERCISE 2: Step 10

source	term name Gene Ontology (Biological process)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
BP	negative regulation of cellular component organization	GO:0051129	385	392	22	3.44e-02
BP	regulation of microvillus assembly	GO:0032534	4	392	3	2.23e-02
BP	positive regulation of cell death	GO:0010942	423	392	23	5.00e-02
BP	regulation of protein kinase B signaling	GO:0051896	82	392	9	4.76e-02
BP	positive regulation of protein kinase B signaling	GO:0051897	56	392	9	2.06e-03
source	term name Biological pathways (KEGG)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
ke	TGF-beta signaling pathway	KEGG:04350	79	386	8	5.00e-02
source	term name Biological pathways (Reactome)	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
re	Regulation of mitotic cell cycle	REAC:453276	85	387	9	5.00e-02
re	APC/C-mediated degradation of cell cycle proteins	REAC:174143	85	387	9	5.00e-02